Exhibit 2

Message
_____

From: Laurens van der Maaten (Google Docs) [comments-noreply@docs.google.com]
Sent: 10/18/2023 2:08:32 AM
To: Nikolay Bashlyko█████████@meta.com]
Subject: Observations on LibGen-SciMag

New activity in the following document

Observations on LibGen-SciMag

For you
Comments

Observations

Shruti Bhosale
• 2:51 PM, Oct 16 (PDT)

cc: ███████meta.com Laurens had some interesting observations about scimag

Nikolay Bashlykov
• 2:53 AM, Oct 17 (PDT)

Thanks for the comments ███████@meta.com!

1 author reacted with 👍 (Laurens van der Maaten)

Laurens van der Maaten
• 7:00 PM, Oct 17 (PDT)
New

Excited to see all the amazing work you and the team have doing on LibGen, ███████@meta.com! The data looks really good :)

ReplyOpen

5 comments

author indexes or whether it is better to try and remove them

Nikolay Bashlykov
• 7:07 AM, Oct 17 (PDT)

we didn't remove the authors in Galactica project, but yes we can experiment on that as well

Laurens van der Maaten
• 7:01 PM, Oct 17 (PDT)

New

Probably also not a p0. Just something I was wondering about.

ReplyOpen

some ablations on different ways to encode references

Nikolay Bashlykov
• 8:06 AM, Oct 17 (PDT)

Discussed with ███████@meta.com that in Galactica the main focus for this format of the references ([START_REF]) was to have the ability to predict the paper based on the context. Not sure we need this ability for our general purpose model.

I was thinking of removing the references completely, but then you could end up with incomplete sentences like this:
"As [2] suggests, ..." -> "As suggests". So I decided to leave the references as is for now.

If we have cycles to ablate more that, would be interesting to see the outcomes.

Laurens van der Maaten
• 6:59 PM, Oct 17 (PDT)

New

Got it, okay. If we've thought about this and made an intentional decision to keep it as-is, that sounds good to me. It would be interesting to do an ablation on this at some point but I agree it is probably not a p0.

ReplyOpen

math hasn't been converted to Markdown

Nikolay Bashlykov

• 8:21 AM, Oct 17 (PDT)

I think LaTeX is better for math. But yeah, we can change it if needed

Laurens van der Maaten
• 7:02 PM, Oct 17 (PDT)
New

How did we decide what stuff we convert to Markdown and what we keep in LaTeX?

And perhaps more importantly: Are we doing this consistently across all data sources? That kind-of seems important (but I may be wrong!).

ReplyOpen

Acknowledgements sections or can we remove them?

Nikolay Bashlykov
• 8:26 AM, Oct 17 (PDT)

it's usually some good text there, but I don't have strong opinion on that. Unless there is PII data there

Laurens van der Maaten
• 7:03 PM, Oct 17 (PDT)
New

It's usually a list of funding agencies, project numbers, and people who proofread the paper, right? Doesn't seem super-valuable but I may be wrong..?

ReplyOpen

---

remove more copyright

Nikolay Bashlykov
• 3:07 AM, Oct 17 (PDT)

yes, agree, it shouldn't be there. Currently we are removing all the copyright paragraphs from beginning and the end of the paper.
But we can also try to remove any line containing "ISBN", "Copyright", "©", "All rights reserved", ...

Laurens van der Maaten
• 7:01 PM, Oct 17 (PDT)
New

👍

I don't know how easy or hard this is but it seems to me like it should be doable for most papers and give us a nice little boost.

ReplyOpen

Google LLC, 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA

You have received this email because you are a participant in the updated items or subscribed to notifications on Observations on LibGen-SciMag. Change what Google sends you. You cannot reply to this email.